3M

УДК 004.89, 004.93

Мурыгин К.В.
Институт проблем искусственного интеллекта МОН и НАН Украины, г. Донецк
kir@iai.donetsk.ua

Концепция системы распознавания речи на основе чтения по губам

В статье рассматривается проблема построения автоматической системы чтения с губ на основе интеллектуального анализа видеоизображений лица диктора. Разрабатываемая система предназначена для обучения пользователя навыкам правильной артикуляции для упрощения визуального восприятия украинской речи людьми с нарушениями слуха и заключается в контроле правильности произнесения обучаемым известных слов.

Введение

В настоящее время системы автоматического чтения с губ в большинстве своем используются для дополнения звукового информационного канала визуальным, что необходимо для повышения качества распознавания речи в условиях шума или посторонних источников звука. Проведенный анализ современного состояния в задаче автоматизации чтения с губ показал, что достигаемые результаты при использовании только визуальной информации являются не вполне удовлетворительными. Объясняется это ограниченностью орального алфавита (алфавита визем – зрительного аналога фонем), что не дает возможности полного описания фонетической структуры языка соответствующими визуальными образами. Практические данные о возможности чтения с губ подготовленными людьми объясняются возможностью применения знания контекста и смыслового комбинирования, которые обеспечивают компенсацию недостатков сокращенного алфавита визем. При автоматизации процесса чтения с губ такую компенсацию можно получить с использованием автоматического семантического анализа, что в настоящее время не является вполне осуществимым. Поэтому в качестве основной, практически достижимой цели можно выделить создание компьютерной информационной технологии для обучения правильной артикуляции при произнесении украинской речи. Разрабатываемая система предназначена для обучения пользователя навыкам правильной артикуляции для упрощения визуального восприятия украинской речи людьми с нарушениями слуха и заключается в контроле правильности произнесения обучаемым известных слов.

Формирование словаря визем

Согласно фонетике украинскому языку присущи 6 гласных и 32 согласных фонемы:

[і], [и], [е], [у], [о], [а];

[б], [п], [д], [д'], [т], [т'], [ґ], [к], [ф], [ж], [з], [з'], [ш], [с], [с'], [г], [х], [дж], [дз], [дз'], [ч], [ц], [ц'], [в], [й], [м], [н], [н'], [л], [л'], [р], [р'].

Здесь *' – означает мягкий звук *.

Сопоставляя фонетический состав украинского языка с исследованиями В.И. Бельтюкова для русского языка [1] и учитывая фонетические сходства украинского и русского языков, можно сформировать следующий оральный алфавит украинских звуков (визем).

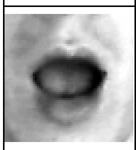
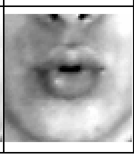
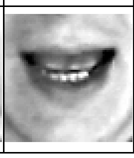
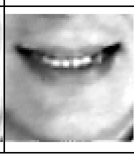
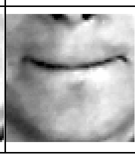
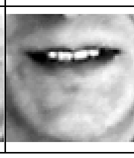
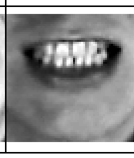
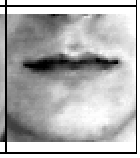
Таблица 1 – Оральный алфавит украинских звуков (визем), полученный по аналогии с алфавитом В.И. Бельтюкова

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
а	о	у	е	і	п	ф	ш	л'	р	с'	т	т'	к	й
				и	б	в	ж	р'		с	д	д'	г	
					м		ч			з'	н	н'	х	
							дж			з	л		ґ	
										ц				
										ц'				
										дз				
										дз'				

Предварительный анализ возможности автоматической классификации образов такого алфавита показал необходимость его существенного сокращения в направлении использования базовых или опорных визем [1], [2]. В приведенной табл. 1 виземы начиная с девятой являются плохо различимыми даже человеком с его значительно более мощным зрительным аппаратом. Это во многом связано с тем, что процесс воспроизведения соответствующих им звуков в значительной мере скрыт внутри ротовой полости, что существенно усложняет их зрительное восприятие и тем более автоматическое распознавание на основе полученного цифрового изображения.

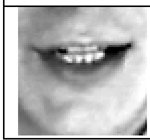
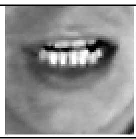
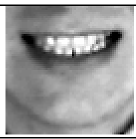
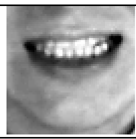
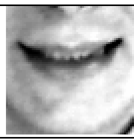
Поэтому для дальнейших исследований в направлении разработки системы автоматического чтения с губ можно принять следующий рабочий алфавит визем, за основу которого приняты опорные виземы (табл. 2).

Таблица 2 – Рабочий алфавит визем

1	2	3	4	5	6	7	8	9
а	о	у	е	і	п	ф	ш	$
				и	б	в	ж	
					м		ч	
							дж	
								

Для сравнения с более широким алфавитом (табл. 1) в табл. 3 приведены также визуальные образы визем для элементов алфавита, не вошедших в рабочий алфавит (табл. 2).

Таблица 3 – Элементы расширенного алфавита, не вошедшие в рабочий алфавит

9	10	11	12	13	14	15
л'	р	с'	т	т'	к	й
р'		с	д	д'	г	
		з'	н	н'	х	
		з	л		ґ	
		ц				
		ц'				
		дз				
		дз'				
						

Как видно из табл. 3, приведенные в ней элементы алфавита визуально трудноразличимы с элементами рабочего алфавита, что может существенно затруднить распознавание произнесенного звука по изображению соответствующей конфигурации губ. Так виземы 10 и 12 визуально трудноотличимы от 8, а виземы 9, 11, 13 и 15 легко спутать как между собой, так и с виземой 5 принятого рабочего алфавита. Знак «$» в рабочем алфавите визем означает нормальное положение, молчание, паузу или любую другую визему, не входящую в этот алфавит. Таким образом, при распознавании предпочтение отдается виземам 1 – 8, а в случае отказа от распознавания – не распознана ни одна из восьми – данной конфигурации приписывается значение 9, которое также может генерироваться в случае промежуточного положения между двумя и более виземами.

Последовательный анализ видеоданных

Для реализации распознавания артикуляции при произнесении речи последовательный анализ видеоданных содержит следующие основные этапы:

– поиск области лица – наиболее перспективным является подход, основанный на использовании: интегрального изображения, каскадного механизма классификации, метода AdaBoost для обучения классификации;

– поиск области губ – может быть решен аналогично поиску лица, а также с использованием активных или гибких контуров;

– распознавание визем – для решения возможно использование многоклассового AdaBoost-метода или объединения двухклассовых классификаторов на основе принципа дихотомии, анализа формы контуров губ, алгебраического подхода, главных компонент.

Основная сложность при решении последней задачи заключается в существенном влиянии на изображения области губ таких плохо контролируемых факторов, как условия освещения и индивидуальные особенности лиц. Для их успешного учета при распознавании необходимо иметь достаточно обширную выборку изображений, отражающую возможные влияния приведенных факторов. В противном случае необходимо вводить соответствующие ограничения на условия эксплуатации системы обучения.

Обнаружение и принцип дихотомии

Построению практически любой системы автоматического распознавания объектов предшествует этап их обнаружения, за которым выполняются дальнейшие действия, включающие извлечение признаков, применение классификаторов и принятие решения о принадлежности. Это особенно актуально для систем распознавания зрительных образов, когда необходимо работать с двумерными данными, представляющими собой отображение трехмерных объектов. В этом случае предварительный этап обнаружения позволяет как повысить качество извлечения признаков, то есть инвариантность к факторам смещения и масштаба, так и скорость выполнения этой операции за счет применения только к выделенной части изображения, как правило, значительно меньшей, чем само изображение. Кроме этого, в ряде практических задач само обнаружение объекта может являться конечной целью анализа изображения, что справедливо для различного рода систем видеонаблюдения.

При достижении высоких показателей работы систем обнаружения объектов, соответствующие методы обучения распознаванию образов могут быть успешно применены и к решению задачи классификации. Это означает переход от задачи классификации двух классов (объект / не объект) к мультиклассовой задаче. В случае если множество распознаваемых классов известно и жестко задано, что справедливо для рассматриваемой задачи распознавания визем, то переход от решения двухклассовой к решению мультиклассовой задачи классификации осуществим на основе принципа дихотомии – представлении многоклассового классификатора в виде последовательности двухклассовых. При этом для достижения высоких показателей быстродействия первые двухклассовые классификаторы достаточно обучить на классификацию объектов, соответствующих классам с наибольшими априорными вероятностями. В задаче автоматического чтения с губ получить оценки априорных вероятностей классов визем, составляющих рабочий алфавит, можно путем статистического определения частот встречаемости каждой виземы в вербальной информации заданной предметной области.

Накопление и состав обучающей базы данных

Исходя из целей обучающей системы – выработка правильной артикуляции при произнесении украинской речи – для разработки алгоритма распознавания артикуляции и оценки правильного произношения необходима репрезентативная выборка примеров правильного произношения фонем украинского языка в виде изображений отдельных визуальных частиц речи согласно используемому оральному алфавиту. Получить такую базу данных достаточно сложно ввиду отсутствия необходимого количества людей уже владеющих правильной артикуляцией и средств проверки их навыков. С другой стороны, как отмечается в ряде работ, в том числе в работе Ф.Ф. Рау [3], при обучении чтению с губ людей педагогам не ставится задача использовать утрированную ярко выраженную артикуляцию. Основными требованиями к произношению в процессе обучения являются замедленный темп речи, подчеркивание ритмико-интонационной стороны речи, соблюдение правил орфоэпии. Исходя из этого при формировании обучающей базы данных алфавита визем можно использовать следующую методику. Испытуемому предлагается изображение правильного произнесения виземы, после чего он повторяет это произнесение. Результат контролируется оператором и сохраняется в виде изображения. Такая последовательность повторяется для каждой виземы принятого алфавита и в различных условиях съемки, таких, как освещение и расстояние до камеры. В результате формируется база данных визем, учитывающая индивидуальные особенности артикуляционного аппарата каждого испытуемого, возможные изменения освещения лица и его масштаба.

Ограничения на скорость обработки видеопотока и отслеживание лиц

По данным исследований в области систем распознавания речевой информации по аудиоданным временной интервал, в котором фонемы можно приблизительно считать стационарными, составляет около 10 миллисекунд. Отсюда скорость захвата и обработки речевой (как аудио, так и видео) информации должна в лучшем случае составлять не менее 100 информационных квантов в секунду. Для неспециализированных устройств захвата видео этот показатель может достигать до 30 кадров в секунду, а выполняемая обработка может снизить скорость потока данных еще больше. Таким образом, используемые методы обработки видеоданных должны быть вычислительно не сложными, позволяющими обрабатывать данные в потоке со скоростью не ниже 25 кадров в секунду. Наиболее вычислительно сложным методом в рассматриваемой задаче является метод обнаружения лица, входными данными для которого является всё изображение [4], [5]. В отличие от него, например, для метода распознавания входными данными является небольшая область изображения, признанная областью губ. Для ускорения работы метода обнаружения лица наряду с каскадным методом классификации возможно использование механизма слежения за лицом (face tracking). Согласно этому механизму положение лица в потоке обрабатываемых кадров ищется один раз – на первом кадре, после этого его положение лишь корректируется путем поиска лица в некоторой, небольшой по сравнению со всем кадром, области, вокруг найденного положения на предыдущем кадре (рис. 1). Размеры этой области должны соответствовать допустимым изменениям положения лица в кадре с учетом скорости его возможного перемещения и частоты захвата нового кадра. В результате использования механизма слежения общая скорость обработки видеоданных возрастает.

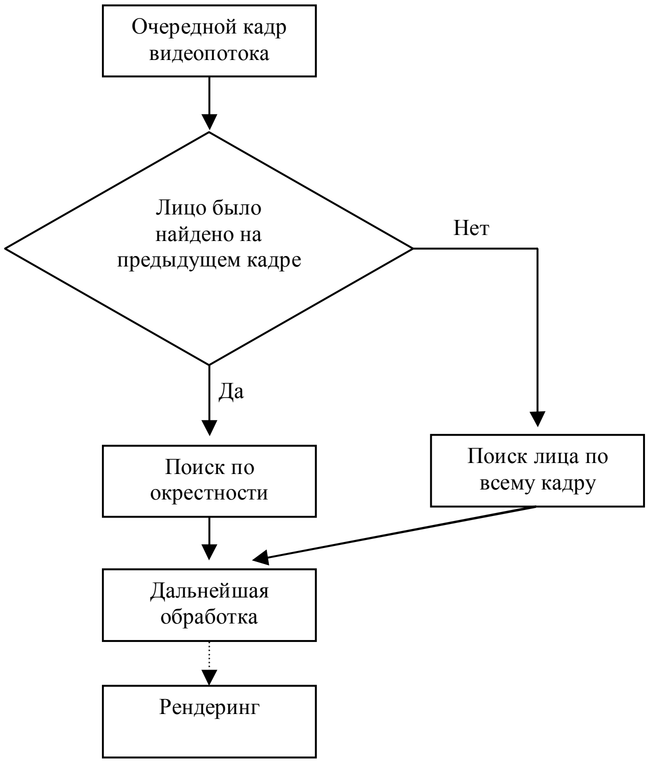


Рисунок 1 – Схема механизма слежения за лицом

Каскадная классификация

Процесс классификации с применением каскада классификаторов заключается в отсеве как можно большего числа ложных изображений объекта на каждом этапе каскада при условии заданной максимальной ошибки пропуска цели (рис. 2).

На каждом этапе итоговый классификатор может быть сформирован согласно методу обучения AdaBoost [6] или любому другому методу обучения классификации, на основе заданных ограничений на ошибку ложного обнаружения и ошибку пропуска цели. При этом следует учитывать, что, как правило, число кандидатов на изображение объекта много больше, чем истинных изображений объектов. Общее число кандидатов при анализе изображения размером $W \times H$ и M возможных масштабов объекта равно:

$$N = \sum_{i=0}^{M-1} N_i = \sum_{i=0}^{M-1} (W - w_i) \cdot (H - h_i) = \sum_{i=0}^{M-1} (W - w_0 k^i) \cdot (H - h_0 k^i),$$

где k – масштабный коэффициент, определяющий отношение между соседними масштабами. Откуда для изображения размером 640*480 и 10 масштабов объекта, число кандидатов $N \approx 3000000$, среди которых правильным может являться только 1. Отсюда следует, что для правильного анализа изображений такого размера необходимо, чтобы после прохождения всего каскада классификаторов ошибка ложного обнаружения должна быть порядка 10^{-6} – 10^{-7} при достаточно низкой ошибке пропуска цели, для которой приемлемыми значениями являются значения меньшие 0,05.

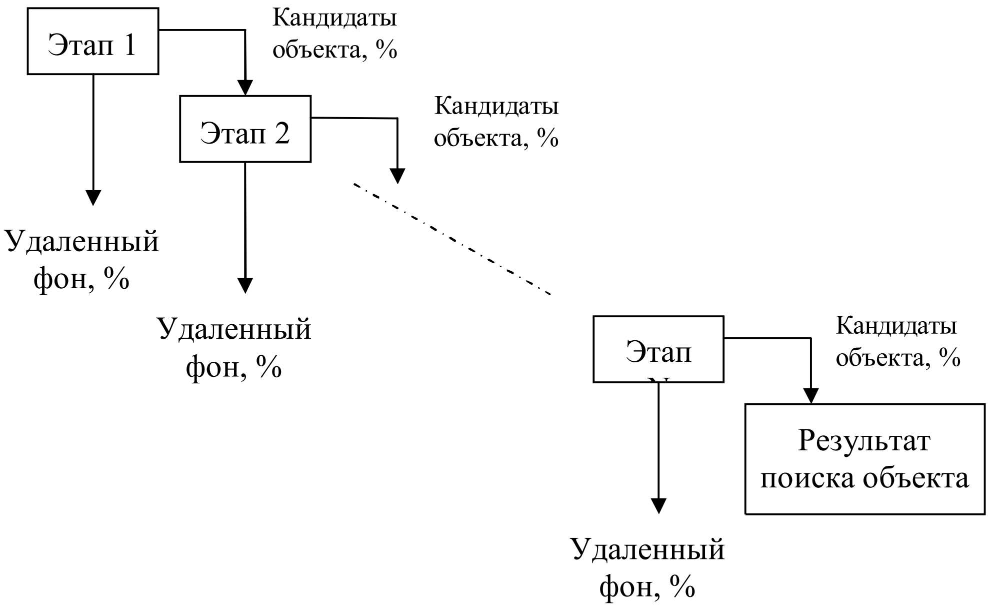


Рисунок 2 – Схема каскадной классификации

Схема обучения по AdaBoost

Схема обучения по AdaBoost представляется достаточно мощным инструментом для решения задачи распознавания, и в сочетании с использованием элементарных (или простых) классификаторов в виде прямоугольных свойств, является достаточно удобной для использования в области автоматического анализа изображений, поиска объектов на изображении, распознавания изображений.

Согласно методу AdaBoost [4], на каждом этапе обучения отбирается элементарный классификатор, дающий минимальную ошибку на текущей базе данных. После этого обучающая база изображений перевзвешивается таким образом, что веса правильно классифицированных изображений уменьшаются, а веса ошибочно классифицируемых экземпляров увеличиваются. Таким образом, на следующем витке обучения поиск наилучшего элементарного классификатора будет в значительной степени зависеть от результатов работы предыдущих отобранных классификаторов, а новый найденный лучший элементарный классификатор будет в большей степени направлен на классификацию плохо разделенных изображений объекта и фона на предыдущих этапах.

Поиск лучшего элементарного классификатора на каждом этапе представляет вычислительно очень сложную задачу. Это объясняется, во-первых, необходимостью использовать достаточно обширную базу данных изображений объекта и фона для их наиболее полного признакового описания и, следовательно, возможности надежного разделения. Во-вторых, большим количеством самих элементарных классификаторов, для каждого из которых необходимо по имеющейся взвешенной базе данных определить соответствующую ему ошибку классификации. Совокупность этих двух особенностей приводит к выводу о том, что для отбора наилучшего признака необходим более эффективный метод, чем полный перебор. Как показали проведенные предварительные исследования, приемлемой скорости обучения можно достигнуть на основе поиска признаков на каждом этапе с применением комбинации метода статистических испытаний и метода градиентного спуска.

Выводы

В статье приводится концепция создания экспериментальной технологии распознавания речи по губам, которая явилась результатом всестороннего анализа современного состояния проблемы автоматического чтения с губ. В ходе проведенной декомпозиции задачи выделены три основных этапа анализа входных видеоданных: поиск лица на изображении, поиск области губ, идентификация конфигурации губ. Рассмотрены основные сложности, возникающие при построении системы, а также пути их возможного разрешения. На основе проведенного анализа литературных источников сформирован рабочий алфавит визуальных образов речи (визем) и выработаны основные принципы построения системы автоматического чтения с губ. Согласно выработанной концепции дальнейшими направлениями исследований могут быть решения задач трех выделенных этапов обработки входных данных. Их успешное разрешение позволит создать экспериментальную технологию автоматического чтения по губам, которая позволит улучшить параметры систем звукового распознавания речи в условиях шума или нескольких дикторов, а также разработать прикладную программу обучения правильной артикуляции для облегчения понимания речи по губам людьми с нарушениями слуха.

Литература

1. Миронова Э.В. Оценка навыка чтения с губ / Э.В. Миронова. – М. : Педагогика, 1980.
2. Режим доступа : http://www.pedlib.hut.ru/Books/pravdina/pravdinap124.html.
3. Методика обучения глухих устной речи: учеб. пособие для студентов дефектол. фак. фед. ин-тов / [ред. проф. Ф.Ф. Рау]. – М. : Просвещение, 1976. – 279 с.

4. Мурыгин К. В. Поиск области лица на изображении методом сопоставления с эталоном с использованием нескольких шаблонов / Кирилл Владимирович Мурыгин // Проблемы бионики. – 2003. – № 59. – С. 55-59.
5. Мурыгин К.В. Автоматический анализ цифровых изображений с целью обнаружения лиц с боковыми поворотами / Кирилл Владимирович Мурыгин // Искусственный интеллект. – 2005. – № 3. – С. 649-657.
6. Paul Viola. Robust real-time object detection / Paul Viola and Michael J. Jones // Proc. of IEEE Workshop on Statistical and Computational Theories of Vision. – 2001.

К.В. Муригін

Концепція системи розпізнавання мови на основі читання по губах

У статті розглядається проблема побудови автоматичної системи читання з губ на основі інтелектуального аналізу відеозображення обличчя диктора. Розроблювана система призначена для навчання користувача навиками правильної артикуляції для спрощення візуального сприйняття української мови людьми з порушеннями слуху і полягає у контролі правильності вимови научуваних відомих слів.

K.V. Murygin

Concept of Speech Recognition Based on Lip Reading

The article is devoted to the concept of development of speech recognition experimental technology on the basis of lip reading. The concept which has been developed is a result of overwhelming analysis of a modern view on the problem of automatic lip reading. In accordance with performed decomposition of the common task the following three principal stages of analysis of entry video data have been determined: search for the face on the image, detection of lips area, identification of lips configuration. The main complications of the system development and methods of their probable solution are given in the article. Having analyzed the sources the working alphabet of visual patterns of speech (visem) and main principles of creating the system of automatic lip reading have been developed. According to the worked out concept further directions of researches can be connected to solving of three selected stages of processing of input information. Their successful solving will allow creating experimental technology of automatic lip reading which will make it possible to improve productivity of speech recognition systems based on audio information channel in the conditions of noise or several speakers, and also to develop a learning application of correct articulation to make speech comprehension based on lip reading by deaf people or people with bad hearing easier.

Статья поступила в редакцию 11.02.2009.